RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Geosciences Research Data Repositories Landscape: re3data.org as a datasource

Alexandre Ribas Semeler
Universidade Federal Rio Grande do Sul, Instituto de Geociências, Porto Alegre, RS, Brasil
ORCID: https://orcid.org/0000-0002-8036-4271
alexandre.semeler@ufrgs.br

Luana Sales
Instituto Brasileiro de Informação em Ciência e Tecnologia, Programa de Pós-graduação em Ciência da Informação, Rio de Janeiro, RJ, Brasil
ORCID: https://orcid.org/0000-0002-3614-2356
luanasales@ibict.br

Adilson Luiz Pinto
Universidade Federal de Santa Catarina, Programa de Pós-graduação em Ciência da Informação, Florianópolis, SC, Brasil
ORCID:https://orcid.org/0000-0002-4142-2061
adilson.pinto@ufsc.br

Carlos Luis González-Valiente
European Alliance for Innovation, Bratislava, Eslováquia
ORCID: https://orcid.org/0000-0002-1836-5257
carlos.valiente89@gmail.com

ARTIGOS

DOI: https://doi.org/10.26512/rici.v17.n3.2024.53645

Recebido/Recibido/Received: 2024-04-28
Aceito/Aceptado/Accepted: 2024-08-06
Publicado/Publicado/Published: 2024-11-10

Abstract
The so-called research data repositories are an evolution of document repositories that aim to access and preserve all research materials used before, during, and after scientific research. Against this backdrop, this study takes an exploratory and descriptive approach to the international Geosciences Research data Repositories (RDR-GEO)landscape, castrated and available on the Registry of Research Data Repositories (re3data.org). Specifically, the metadata related to themes, countries, licenses, software, institutions, and research data used by the international Geosciences community is analyzed. The data was extracted from re3data.org using web scraping techniques. Once the data has been collected, it is necessary to use the OpenRefine software to clean, recode, merge data sets, and export data for visualization in tables, graphs, and network maps. It is concluded that although the international community of data librarians in Geosciences encourages using software such as Ckan, DSpace, DataVerse, e-Prints, and Fedora, they are little used in Geosciences and mostly use their technologies. Between the (325) RDR-GEO, only (17) use them. Some examples of RDR-GEO made with this software are GeoPlatform (Ckan), DataShare (DSpace), UCLA Social Science Data Archive (Dataverse), OceanRep GEOMAR Repository (e-Prints) e, and TAMBORA (Fedora). Finally, RDR-GEOs are heterogeneous data sources that enable access and preservation of

various research data types. They store information/knowledge of areas related to Earth Sciences such as Atmospheric, Geodesic, Geophysical, Geological and Oceanographic Sciences.
Keywords: Data Repository. Research Data. Geosciences. Re3data.

Repositórios de dados de pesquisa no domínio das Geociências: re3data.org como fonte de dados

Resumo

Os chamados repositórios de dados de pesquisa são uma evolução dos repositórios de documentos e visam aceder e preservar todos os materiais de pesquisa utilizados antes, durante e após a investigação científica. Com base nesse contexto, este estudo realiza uma abordagem exploratória e descritiva do cenário internacional dos Repositórios de Dados de Pesquisa em Geociências (RDP-GEO) castrados e disponíveis no *Registry of Research Data Repositories* (re3data.org). Especificamente, são analisados os metadados relacionados com os temas, países, licenças, software, instituições e tipos de dados de pesquisa utilizados pela comunidade internacional de Geociências. Os dados foram extraídos do re3data por meio de técnicas de web scraping.Uma vez coletados os dados, é necessário utilizar o software OpenRefine para limpar, recodificar, fundir conjuntos de dados e exportar dados para visualização em tabelas, gráficos e mapas de rede. Conclui-se que, embora a comunidade internacional de bibliotecários de dados em geociências incentive o uso de softwares como Ckan, DSpace, DataVerse, e-Prints e Fedora, eles são pouco utilizados na área de Geociências que utiliza outras tecnologias ligadas a geração dos dados, entre as (325) repositórios, apenas (17) as utilizam. Alguns exemplos de RDR-GEO feitos com este software são GeoPlatform (Ckan), DataShare (DSpace), UCLA Social Science Data Archive (Dataverse), OceanRep GEOMAR Repository (e-Prints) e, e TAMBORA (Fedora). Por último, os RDP-GEOs são fontes de dados heterogéneas que permitem o acesso e a preservação de vários tipos de dados de pesquisa e seus metadadoscontemplando áreas como as Ciências Atmosféricas, Geodésicas, Geofísicas, Geológicas e Oceanográficas.
Palavras-chave: Repositório de dados. Dados de pesquisa. Geociências. Re3data.

Panorama de los repositorios de datos de investigación en Geociencias: re3data.org como fuente de datos

Resumen

Los denominados repositorios de datos de investigación son una evolución de los repositorios de documentos y tienen como objetivo acceder y preservar todos los materiales de investigación utilizados antes, durante y después de la investigación científica. En este contexto, el presente estudio adopta un enfoque exploratorio y descriptivo del panorama internacional de los Repositorios de Datos de Investigación en Geociencias (RDI-GEO) castrados y disponibles en el *Registry of Research Data Repositories* (re3data.org). En concreto, se analizan los metadatos relacionados con los temas, países, licencias, software, instituciones y tipos de datos de investigación utilizados por la comunidad internacional de Geociencias. Los datos se extrajeron de re3data mediante técnicas de web scraping. Una vez recopilados los datos, es necesario utilizar el software OpenRefine para limpiarlos, recodificarlos, fusionar conjuntos de datos y exportarlos para su visualización en tablas, gráficos y mapas de red. Se concluye que, aunque la comunidad internacional de bibliotecarios de datos geo científicos fomenta el uso de software como Ckan, DSpace, DataVerse, e-Prints y Fedora, son poco utilizados en el área de las geociencias que más utiliza sus tecnologías. De los (325) RDR-GEO, sólo 17 los utilizan. Algunos ejemplos de RDR-GEO realizados con este software son GeoPlatform (Ckan), DataShare (DSpace), UCLA Social Science Data Archive (Dataverse), OceanRep GEOMAR Repository (e-Prints) y TAMBORA (Fedora). Por último, los RDI-GEOs son fuentes de datos heterogéneas que permiten el acceso y la preservación de diversos tipos de datos de investigación almacenan información/conocimientos de áreas relacionadas con las Ciencias de la Tierra, como las Ciencias Atmosféricas, Geodésicas, Geofísicas, Geológicas y Oceanográficas.
Palabras clave: Repositorio de datos. Datos de investigación. Geociencias. Re3data.

1 Introduction

The multitude of technological systems designed to facilitate research data management constitute a rich semantic-relational universe of information and knowledge, which documents the outcomes of the scientific research process. The diversity of pathways through which research data flows, particularly within the web context, highlights the imperative to organize, comprehend, conserve, and analyze the information and knowledge that can be derived from this digital medium.

Digital document repositories emerged in the early 1990s to disseminate publications and other types of digital objects. Developed using free software, they are linked to movements to open knowledge, such as the Open-Source Initiative, Open Access, Open Data, and Open Science. One of the first initiatives to create repositories was arXiv, developed by Cornell University in the USA. It began in 1991 as a digital library for preprints in physics (Rice; Southall, 2016).

Digital data repositories, including research data repositories (RDR), facilitate data sharing and contribute significantly to the open access and open science movement (Benin; Hamanaka; Gonçalez, 2022). According to Valles-Coral *et al*. (2023), these repositories provide the opportunity to establish interoperability features, encourage collaboration, and exchange information between different scientific entities.

The emergence of research data repositories is currently being discussed. Technological and organizational information systems help researchers manage and digitally curate their research data. According to Panduro (2023), the evolution of research data management tools, especially in dynamic and contingent data management, explores the adoption and implementation of technologies.

The global distribution of research data repositories is cataloged by the Registry of Research Data Repositories (re3data), the international registry of research data repositories. Founded by the German Research Foundation (DFG), re3data brings together initiatives from the Library and Information Services (LIS) of the GFZ German Research Center for Geosciences, the Library of the Karlsruhe Institute of Technology (KIT), the School of Library and Information Science (BSLIS) at the Humboldt-Universität in Berlin and the Libraries of the Purdue University in Germany. Since the fall of 2012, re3data.org has been internationally indexing some (3198, March 2024) research data repositories. It offers researchers, funding organizations, libraries, and publishers a systematic overview of the heterogeneous landscape of research data repositories (Pampel *et al*., 2023; Re3data, 2024).

We chose the Geosciences area because it has excellent potential for generating data. Fields such as atmospheric sciences, geodesy, geophysics, geochemistry, geology, oceanography, and paleontology produce research data in various formats and types, as well as

analogical and digital forms. Nowadays, Geosciences research data repositories (RDR-GEO) make it possible to access and preserve a wide range of types of research data.

In the context of Geosciences, this means that researchers from different disciplines within Geosciences can share and collaborate on data, leading to a more comprehensive understanding of the Earth's systems. Research data repositories are not just an evolution of documentary repositories; they play a crucial role in Geosciences. Their primary function is to facilitate the search for and access to research data specific to Geosciences.

According to Uzwyshyn (2016), RDRs store, organize, preserve, and share data for reuse in Geosciences research. They are large database infrastructures set up to manage, share, access, and archive researcher's datasets in Geosciences. They are an essential part of the scientific research cyberinfrastructure in Geosciences and aim to preserve, provide long-term access to, and reuse research data specific to Geosciences. Therefore, they need to be planned and structured from the outset of their implementation as information systems in Geosciences.

The current study aims to examine the RDR listed on re3data.org through the prism of Geosciences. The methodological procedures involve understanding and using programming languages and software for web scraping and data refinement. This approach ensures that the data collected from the repositories is comprehensive and accurate. As an object of investigation, research data repositories are a current research trend in Library and Information Science, and this study aims to contribute to the understanding of their role and impact in Geosciences.

The following text presents a brief review of research data repositories, the methodology employed, and the research problem. It also reflects the international Geosciences research data repositories landscape, and finally the discussions.

2 Background: Research Data Repositories

The so-called research data repositories are an evolution of document repositories that aim to access and preserve all research materials used before, during, and after scientific research. Against this backdrop, this study takes an exploratory and descriptive approach to the international Geosciences research data repositories landscape, castrated and available on the Registry of Research Data Repositories, re3data.org.

Research data repositories, a crucial component of the scientific academic record, are increasingly recognized as a fundamental task for librarians and information science professionals. Their role in ensuring the transparency of scientific research, preserving data, and facilitating the use and reuse, analysis, and re-analysis of data is paramount. This guarantees the

continuous advancement of scientific knowledge, underscoring the significance of your work in Geosciences and data management.

In addition, research data repositories, as defined by Pampel*et al.,* (2013);and Uzwyshyn (2016), represent an advancement over document repositories. Their purpose is to provide access to and ensure the preservation of all research materials utilized throughout the scientific research process. These repositories serve to enhance transparency in scientific research, safeguard data, and facilitate the utilization, reuse, analysis, and reanalysis of research data.

A research data repository can be defined as an infrastructure that facilitates the FAIR (Findable, Accessible, Interoperable, Reusable) research data management (Wilkinson*et al*., 2016). According toPeng, Gross and Edmunds(2022),the repository's proponents, it should facilitate the sharing of research data in a manner compatible with FAIR principles, thereby promoting data discovery and visibility. The primary functions of the RDR are the discovery, identification, and referencing of research data, increasing opportunities for data reuse, enabling the preservation, dissemination, and interoperability of data, facilitating the sharing and visibility of data effectively, and promoting data initiatives such as FAIR.

As Kindling *et al.,* (2017), elucidate the function of RDRs in data quality, demonstrating that quality assurance in RDRs is complex and non-linear. While there are some common patterns, data management approaches are individual and diverse across repositories. For instance, it is noteworthy that the repository provides sufficient recognition of the contribution of data reviewers and certification. Nevertheless, the study does not substantiate the assertion that a repository's certification status is a definitive indicator of the repository's provision of data quality assurance.

According to Pampel*et al.* (2023) the global registry of research data repositories, serving as the world's largest directory of RDRs. It allows scientists, funding agencies, libraries, and data centers to locate, identify, and cite RDRs. Its primary objective is to facilitate the discovery, identification, and referencing of research data repositories by providing an accessible metadata service based on open data.

The following text presents the methodology employed to address the research problem, namely the collection, analysis, and visualization of global registry of research data repositories metadata.

3 Materials and Methods

The research method employed in this paper was designed to conduct an exploratory and descriptive study of the international Geosciences research data repositories landscape

available on re3data.org. This comprehensive approach ensures the validity and reliability of our findings.

The research method will help answer the following questions: how many research data repositories registered with re3data are in the Geosciences? Which are the leading institutions? Which countries and in which languages is the research data available? What is the typology of these repositories? What types of data are deposited? What types of access and licenses are used? What are the themes and areas of knowledge? What type of software? These questions are in addition to other questions that may arise during data interpretation.

The method's development required applying knowledge inherent in the techniques and technologies used for descriptive data analysis, information retrieval, data manipulation, data analysis, and data visualization.

The method's applied result is three Python scripts developed by Semeler*et al.,* (2023), see supplementary files in Table 1, available at (https://doi.org/10.17605/OSF.IO/TQ4JB).

Table 1 - Scripts

Name	Link to script	Description
Supplementary File S1	https://osf.io/evnqc	python script to extract the list of links from re3data API
Supplementary File S2	https://osf.io/2wz4e	python script to extract re3data metadata
Supplementary File S3	https://osf.io/xh7zp	python script to create a graph

Source: Elaborate by Authors (2024)

The methodological procedures require the understanding and using programming languages and software for web scraping and data refinement. Thus, the study is divided into the following stages: data collection on the web (web scraping), data manipulation and descriptive analysis (cleaning, organization, disambiguation treatment, similarities and adaptations to data formats), and visualization (creation of graphs and charts). The practical procedures for collecting and analyzing the corpus of data used in this study are explained below.

3.1 Data Collection

The data collection source is re3data.org, which, as of March 2024, indexes around (3.198) research data repositories, and indexes (325) Geosciences repositories. The metadata with the descriptions retrieved from re3data delimits the international distribution of the repositories. It is possible to select information for descriptive analysis of the data.

Data collection sought to select information on the international composition of these repositories. The application programming interface (API) (http://www.re3data.org/api/<apiidentifier>) available on re3data was used to do this; re3data offers retrieval of all or part of its content via API. Currently, the platform offers a simple open search implementation and a representational state transfer (REST) interface version for data extraction (Re3data, 2024).

The XML schema used by re3data, available at (https://schema.re3data.org/4-0/re3dataV4-0.xsd), to describe the repositories is in version 4.0; according to Pampel*et al.* (2023), the re3data metadata represents the structure of the analyzed data. The standard in XML contains metadata properties on the general scope, content, infrastructure, technical, and quality standards for research data repositories.

Thus, the repositories analyzed in this paper were selected using re3data's REST interface API. The query on re3data resulted in a list of links containing the address of each repository description (https://www.re3data.org + API+V1+ repository + cod_repository), (see Supplementary File S4)available at (https://osf.io/shwa3).

According to Kindling *et al*. (2017); and Khan *et al.* (2024),a re3data metadata entry provides general information about the RDR, such as the repository name, uniform resource locator (URL), disciplinary scope and a descriptive paragraph, information concerning the responsible institutions of an RDR, such as the institution's name, type, location and type of responsibility, legal issues, including access and upload regulations, as well as the availability of policies; and technical aspects such as information concerning supported persistent identifier systems, application programming interfaces or software in use if determinable.

To operationalize the collection of repository descriptions, we used web scraping scripts developed by Semeler *et al.,* (2023), (see Supplementary File S5) available at (https://osf.io/5yvr4). We collected each description from the list of re3data links with them. In practice, the collection involved running a web scraping application on the re3data API server.

The data collected on re3data was web-scraped, making it possible to select essential information about the repositories. After web scraping, the data was copied in (.tsv) format. The data extracted corresponds to the name of the repository, its description, its URL, the name of the software, the type of license, the type of repository, the type of content, the language of the data, the institution responsible, the country, and the subjects of the data made available by the repository.

3. 2 Data handling and analysis

Once the data has been collected, it needs to be manipulated to meet the requirements for preparing it for analysis. At this point, it is necessary to incorporate some data refinement software that can clean, recode, merge datasets, and import and export data from one type of software to another without losing content. Analytical datasets can be generated using OpenRefine, software available at (https://openrefine.org/), which is used to load, clean, reconcile, categorize, and convert data from one format to another.

OpenRefine, known initially as Google Refine, is a data cleaning software that prepares data for pre-analysis. OpenRefine contains a series of clustering and clustering algorithms. It is a software tool used to work with disorganized data. It is used to bring consistency to data. It is not a web service, but a desktop application used to process data. The use of OpenRefine follows the flow:

a) Import the dataset;
b) model the columns for analysis;
c) check the consistency of the data;
d) apply a clustering algorithm to look for words spelled differently, such as "Roma" and "roma". OpenRefine allows the use of the key-collision method and the fingerprint function;
e) generate datasets with the frequency of occurrence of the contents contained in the dataset;
f) convert the dataset for analysis and visualization.

We can gain a deeper understanding of the data by summarizing the data and creating different tables and visualizations after the refinement process. The refinement process enables us to summarize based on frequency description and mine the texts that comprise the dataset collected in re3data.

This leads to the discovery of valuable insights and the ability to perform typical data mining tasks such as classification, clustering, regression, summarization, and association rule learning. These tasks are based on simple tabular data techniques: the rows correspond to the instances, and the columns correspond to the variables. After conducting these research procedures, we summarized the international scenario of Geosciences research data repositories, demonstrating the practical application and impact of the data refinement process.

4 Geosciences Research Data Repositories Landscape

RDRs are sources used to provide access to and preserve research data. They make data available, share it, and provide long-term access. The international composition of these repositories can be mapped using re3data. In total, there are (3.198) research data repositories

registered as of March 2024, of which there are (325) records of cataloged repositories in various areas of knowledge related to the Geosciences, (see Supplementary File S5) available at (https://osf.io/5yvr4), when re3data classifies the area of Geosciences, it includes Geography.

The result of the collection request to re3data helped to answer the initial questions posed in the Materials and Methods of this paper: how many research data repositories registered with re3data are in Geosciences? Which are the leading institutions? Which countries and in which languages is the research data available? What is the typology of these repositories? What types of data are deposited? What types of access and licenses are used? What are the themes and areas of knowledge? What type of software?

4.1 How many research data repositories registered with re3data are in Geosciences? Which are the leading institutions?

The institutions that collect the most Geosciences research data are the GFZ German Research Centre for Geosciences at the Helmholtz Centre Potsdam in Germany, the Woods Hole Oceanographic Institution in the USA, and Environment and Climate Change Canada in Canada.

The GFZ Research Center for Geosciences makes various research data collections available through the Information System and Data Center (ISD), a portal for geoprocessing data services, geospatial coordinate metadata, documentation, and software tools for practicing with geoscientific research data in general. The ISD provides access to collections of data collected by sensors and satellites. One example is the Challenging Minisatellite Payload (CHAMP)[1] data, a German space mission launched in the 2000s to collect geoprocessing data from the Earth's atmosphere.

Another collection from the same center is the International Geodynamics and Earth Tide Service (IGETS), which provides data on the temporal variations of the Earth's gravity field. The center also offers data from the Gravity Recovery and Climate Experiment (GRACE), which is a project between the National Aeronautics and Space Administration (NASA) and the GFZ to generate information on the Earth's gravitational field and measure the planet's water reservoirs in soil, ice, and oceans.

The second institution with the most significant number of repositories in the USA is the Woods Hole Oceanographic Institution. The institute contributes oceanographic datasets, such

[1] Challenging Mini-satellite Payload (CHAMP). Retrieved March 26, 2024, from https://eospso.nasa.gov/missions/challenging-mini-satellite-payload

as the WHOI Ship Data-Grabber System[2], a collection of oceanographic research data gathered from ships underway. The data in this repository is recorded with departure and arrival times, GPS and navigation systems from ships underway. In addition to this collection, the institute makes available the Seafloor Sediments Data Collection, a collection of (14.000) stratigraphic samples and geological sediment sequences recovered from the seabed.

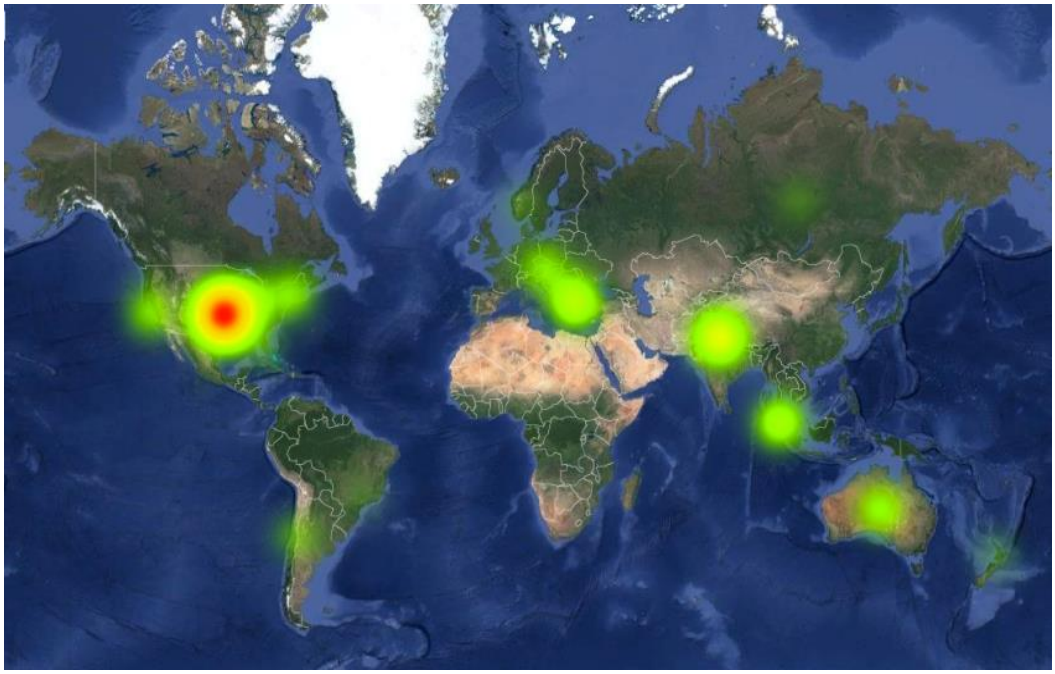
Environment and Climate Change provides repositories in Canada containing data compilations from the National Pollutant Release Inventory. This repository offers collections of research data on releasing pollutants (into the air, water, and land). This data is used to identify pollution prevention priorities and support managing toxic substances and pollutants in the air, water, and land. Another Canadian repository is the Canadian Centre for Climate Modelling and Analysis (CCCMA)[3], which offers research data generated using climate models developed by the CCCMA. This data is used to study climate variability and change in Canada.

4.2 Which countries and in which languages is the research data available?

The following is the answer to the distribution of repositories by countries and institutions with research data collections in the Geosciences, using the minimum occurrence of (2) repositories per institution and country as an analysis factor. To answer the other question, what is the geographical distribution of these repositories, and in what language do they make their research data available? As shown in Fig. 1.

[2] WHOI Ship Data-Grabber System. Retrieved March 26, 2024, from http://4dgeo.whoi.edu/shipdata/index.html.
[3] Canadian Centre for Climate Modelling and Analysis. Retrieved March 26, 2024, from https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/modeling-projections-analysis/centre-modelling-analysis.html

Fig. 1- Geographical distribution of repositories in the Geosciences.



Source: Elaborate by Authors (2024).

By interpreting this descriptive data, we can see that the USA has (264) repositories, making it the country with the most research data repositories. These repositories are distributed among universities, such as Columbia University, Cornell University, and the University of Alaska Fairbanks; among government research departments, such as Goddard Space Flight Center, National Science Foundation, National Aeronautics and Space Administration, Earth Observing System, and Woods Hole Oceanographic Institution.

The overall ranking is divided between the research data repositories of (30) countries, the ones with the most repositories in the area of Geosciences are: USA (123), DEU (53), CAN (32), AAA (16), GBR(13), FRA (12), AUS (11), EEC (11), CHIN (9), JPN (7), IND (6), ITA (5), RUS (5), NOR (3), AUT (2), GRL (2), NLD (2), and (1) repository BEL, CHE, CIV, DNK, ESP, EST, FIN, IDN, KOR, NZL, SAU, SWE, ZAF.

These countries account for around 90% of the repositories. Fig. 1 shows that they are concentrated in North America, Asia, Oceania, and Europe. Concerning the language in which the research data is made available, English is the main language, with only (16) of these repositories in vernacular languages such as German (12), French (2), and only (1) in Spanish and (1) other in Russian.

4.3 What is the typology of these repositories?

Another point to note is the type of repository in Table 2. Research data repositories can be institutional, disciplinary, multidisciplinary, or the result of specific research projects. The main types of repositories are disciplinary and institutional.

Table 2 – List of types of repositories

Disciplinary	228
disciplinary, institutional	51
disciplinary, other	26
Institutional	9
Other	9
disciplinary, institutional, other	2

Source: Elaborate by Authors (2024)

An institution, such as a university or research institution, manages institutional research data repositories. At the university level, their scope is institutional. An example of this type is the UK-based Edinburgh DataShare[4], which makes available datasets produced by the University of Edinburgh. Another type of research data repository presented is disciplinary. One example is PANGAEA - Data Publisher for Earth & Environmental Science[5], which defines itself as an open-access library designed to archive, publish, and distribute georeferenced data on the Earth system. It was developed and maintained by the Alfred Wegener Institute for Polar and Marine Research (AWI) and MARUM[6] - Center for Marine Environmental Sciences at the University of Bremen in Germany. These repositories are aimed at archiving specific research domains.

According to Kindling *at al.* (2017)disciplinary or thematic repositories are extremely varied and heterogeneous, reflecting the multiplicity of disciplines and the diversity of data generated in the context of global scientific research. Alongside disciplinary and institutional approaches, there are multidisciplinary research data repositories. The authors explain that these repositories are those corresponding policies that accept submissions of data collections from various areas of knowledge and from different research institutions. One example cited is

[4]Edinburgh DataShare. Retrieved March 26, 2024, fromhttp://datashare.is.ed.ac.uk
[5] Data Publisher for Earth & Environmental Science. Retrieved March 26, 2024, from http://www.pangaea.de.

[6]Bremen Core Repository. MARUM. Retrieved March 26, 2024, from https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html

FigShare[7], which allows researchers to publish all their data in a citable, searchable, and shareable form. It was developed by Digital Science and Macmillan Publishers Company, an international company based in the United States and the United Kingdom. Finally, there are research repositories focused on research data from specific research projects, such as the Scientific Drilling Database (SDDB)[8], developed by the Scientific Continental Drilling Program (ICDP), making Geosciences datasets available in open access.

An example of a disciplinary repository is ShareGeoOpen, which has been discontinued, so its datasets have been migrated to this Edinburgh DataShare Collection, a spatial research data repository offering free access to maps and geospatial data from data providers in the UK. It was developed as part of the EDINA project and designed to ensure continued access to scientific data by the UK's academic and educational sector.

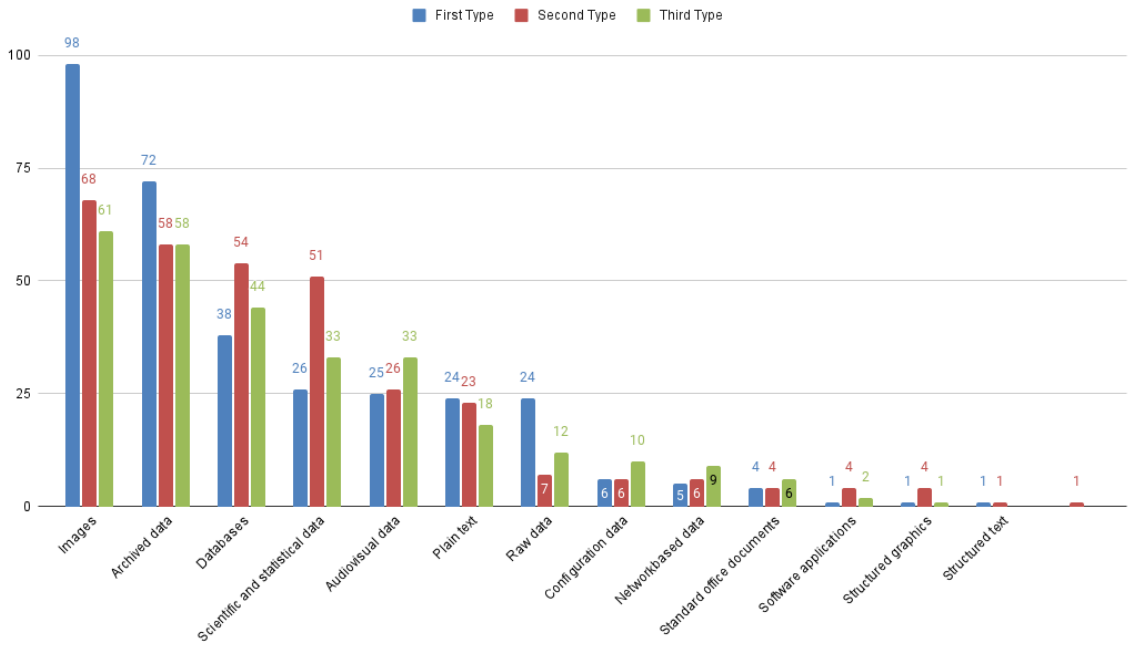
ShareGeo Open is indexed by the Thomson Reuters Data Citation Index. Most of the file formats in this research data collection can be visualized and analyzed using GIS data packages. Edinburgh DataShare is an example of an institutional-type research data repository. It is produced by the University of Edinburgh, Scotland, UK. DataShare is maintained and hosted by the University's Information Services Department. Researchers at the University of Edinburgh who produce research data associated with publications, such as articles or scientific reports, are invited to upload their dataset for sharing and preservation on DataShare, where a persistent identifier is assigned to each dataset deposited.

4.4 What types of data are deposited?

In addition to the type of repositories, it is essential to present the typology of the research data deposited in them. Research data is evidence and input collected, observed, recorded, and created for analysis purposes, and it can produce research results for a scientific study. A research data repository can contain more than one type of research data stored. Graph 1 shows the frequencies of typology of research data in Geosciences described on re3data.org.

[7]Figshare. Retrieved March 26, 2024, from http://figshare.com
[8] Scientific Drilling Database. Retrieved March 26, 2024, from http://www.scientificdrilling.org

Graph 1 - Types of research data



Source: Elaborate by Authors (2024).

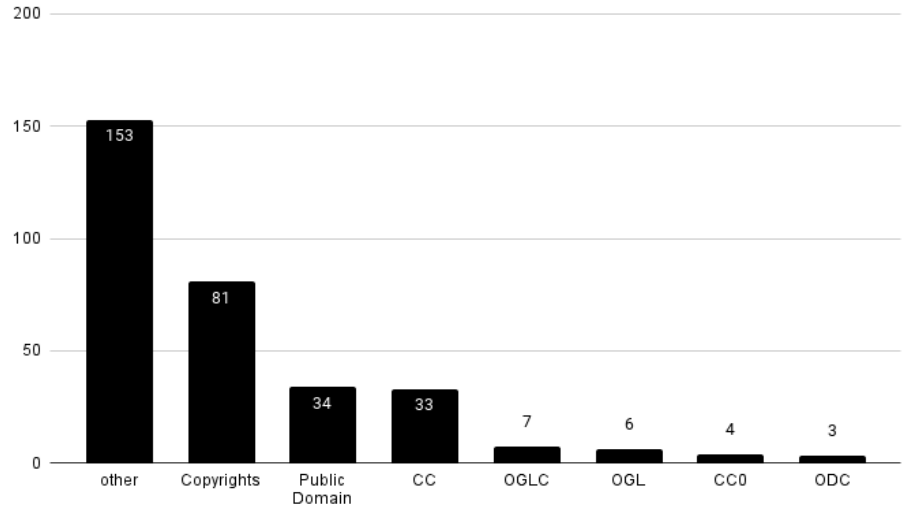
Some repositories make types of data available, such as The Northern California Earthquake Data Center (NCEDC), which disseminates multiple types of data related to earthquake prediction in California, USA. The NCEDC produces seismic data from seismic, geophysical, electromagnetic, water level, and wind speed sensors. Another example is the Cornell University Geospatial Information Repository (CUGIR), [9]which provides geospatial data and metadata on New York State in the USA. However, according to Graph 1, most repositories provide research data of the following types: images, scientific and statistics data raw data, plain text, standard office documents, databases and audiovisual.

4.5 What types of access and licenses are used?

Another question to be answered is what types of access and licenses Geosciences research data repositories use. Most repositories (174) are restricted, (149) are open, and (2) are closed. The type of license used can be seen in Graph 2.

[9] Cornell University Geospatial Information Repository. Retrieved March 26, 2024, from http://cugir.mannlib.cornell.edu.

Graph 2. Types of license data



Source: Elaborate by Authors (2024)

Most repositories (153) use a special license related to the specific use by the institutions offering this data. For example, the Community Data Portal, a disciplinary repository in the USA supported by the University Corporation for Atmospheric Research, although it offers its data openly, requires users to register and accept the terms of use of the data. By registering and accepting the repository's terms, the user receives the right to use the data free of charge for educational and non-commercial purposes.

Another example of open access is Figshare, a repository that allows the publication of research results and adopts Creative Commons Public Domain Dedication (CC0) to release its content globally without restrictions and any reserved rights. Placing research data in the public domain under CC0 removes any legal doubt about the use of the data. Although CC0 does not legally require users to cite the source, it does not affect the ethical standards for attribution in the scientific and research communities. However, most licenses and types of access are specific to each type of data and depend on the institution hosting the repository.

4.6 What are the themes and areas of knowledge?

What follows (Fig. 2) is an attempt to answer which themes and areas of knowledge are present in the repositories. Regarding themes, the repositories are distributed in (36) areas of knowledge subdivided in Table 3, the list of fields of knowledge.

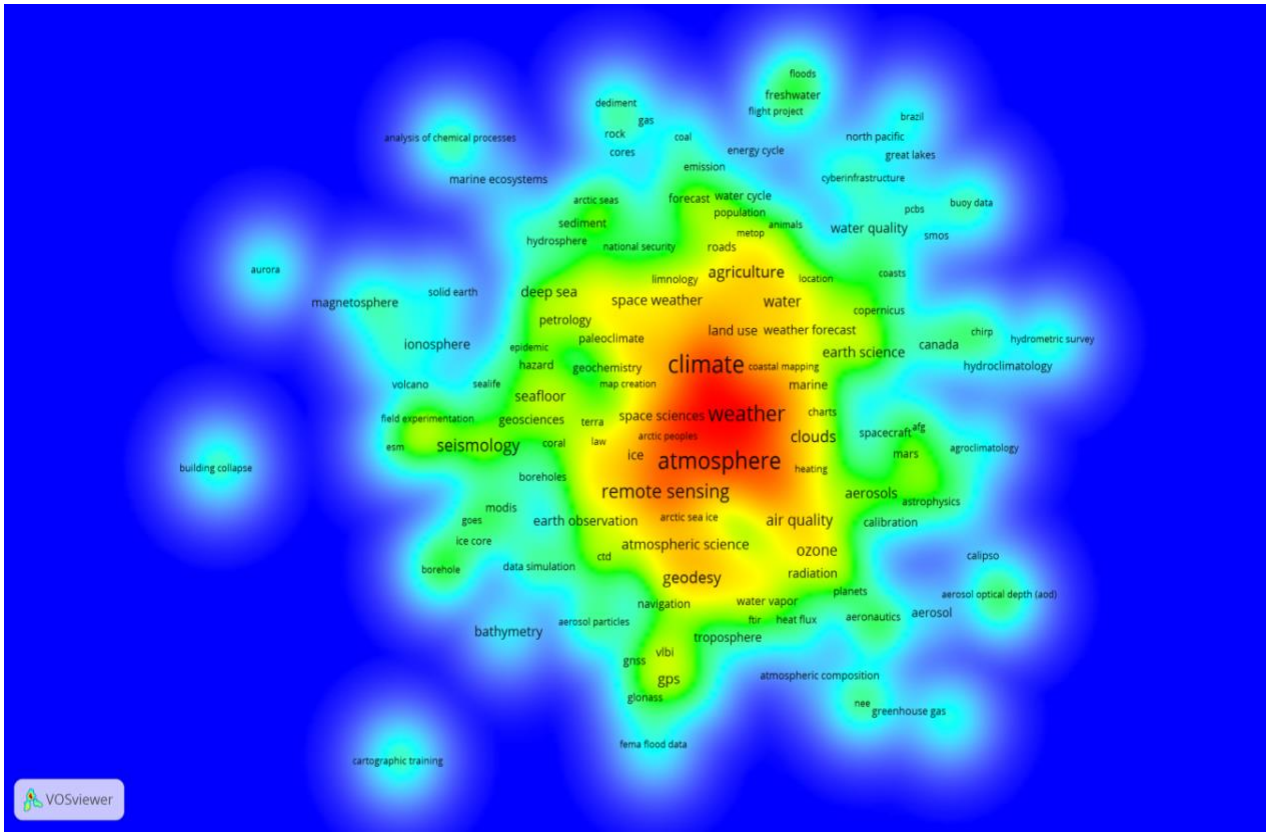
Table 3 – List of fields of knowledge

Fieldsofknowledge	freq.	Fieldsofknowledge	freq.	Fieldsofknowledge	freq.
3 Natural Sciences	311	313 Atmospheric Science and Oceanography	163	34 Geosciences (including Geography)	64
2 Life Sciences	10	315 Geophysics and Geodesy	36	31301 Atmospheric Science	52
1 Humanities and Social Sciences	4	311 Astrophysics and Astronomy	31	315 Geophysics and Geodesy	41
		34 Geosciences (including Geography)	30	31302 Oceanography	38
		314 Geology and Paleontology	12	313 Atmospheric Science andOceanography	21
		318 Water Research	12	31502 Geodesy	18
		317 Geography	9	318 Water Research	10
		202 Plant Sciences	8	Mineralogy and Crystallography	8
		316 Geochemistry	8	31501 Geophysics	7
		31 Chemistry	6	317 Geography	6
		111 Social Sciences	2	20202 Plant Ecology and Ecosystem Analysis	4
		203 Zoology	2	203 Zoology	4
		3 Natural Sciences	2	31101 Astrophysics and Astronomy	4
		304 Analytical Chemistry	2	316 Geochemistry	4
		309 Particles	2	31401 Geology and Paleontology	3
				31801 Hydrogeology	3
				112 Economics	2
				21 Biology	2
				Method Development (Chemistry)	2
				nuclei and Fields	2

Source: Elaborate by Authors (2024).

Fig. 2 shows this distribution of thematic areas according to the keywords used by re3data, the criterion being the minimum occurrence of (2) terms per area.

Fig. 2-Thematic density of repositories in all areas of knowledge



Source: Elaborate by Authors (2024).

According to Fig. 2, which represents the thematic density in research data repositories, there are three main areas. The first area, represented by blue, explores a cluster linked to atmospheric issues and remote sensing. The second area, represented by green, is related to the exact and earth sciences. The third area, represented by yellow, deals with multidisciplinary issues. The density areas have four primary colors (blue, green, yellow, and red). The themes in blue are less recurrent, and in the green band, there are more recurrent issues than in the previous ones, although they are areas of prospection. The yellow zone shows the main themes linked to the cluster's core, which is red. The thematic core, represented by red, includes water, climate, and astrophysics research data.

The research data repositories belonging to the Natural Sciences areas have specific subareas such as (Atmospheric Science, Oceanography, Astrophysics, Astronomy, Geophysics, Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography) and these subareas revealed the most relevant themes in the Geosciences research data.

In this context, it is possible to see (153) themes used to describe RDR-GEOs, among which the most recurrent are meteorology, climatology, geodesy, and oceanography. These themes reflect the main areas of research in the Natural Sciences. The most used keywords in

Natural Sciences reveal the main themes available: telemetry, natural disaster, satellites, site-based (point) ecological data, weather forecast, ice, gas, ocean tides, ocean, seismic data, and earthquakes, among other themes that can be found in the repositories. One example is the World Data Center for Glaciology in Cambridge at the Scott Polar Research Institute. This repository holds collections of research data on snow and ice worldwide, some (40.000) records dating back to 1661.

Life Sciences encompasses (Medicine, Zoology, Biology, Agricultural, Economics, and Sociology), among other sub-areas. The main themes available in Life Sciences are biodiversity, ecology, marine and terrestrial animals, taxonomists, geospatial data, oceanographic data, Arctic, Antarctic, and climatic change, among other themes found in the RDR-GEOs. One example is the Integrated Ocean Drilling Program (IODP) Bremen Core Repository, which provides oceanographic data on the Atlantic and Mediterranean oceans and the Black and the Baltic Seas. The University of Bremen is developing the project and represents German participation in the IODP. The Bremen Core Repository is one of three IODP repositories, such as the Gulf Coast Repository (GCR) in College Station, Texas, USA, and the Kochi Core Center (KCC) in Japan.

Another area of knowledge, Humanities, due to the inclusion of geography as part of Geosciences, has (33) repositories and covers disciplines such as (Ancient Cultures, Prehistory, Education, Sciences, Economics, Social and Behavioural Sciences, Statistics, and Econometrics), (54) terms used to describe repositories in the Humanities. The topics available are multidisciplinary, anthropological, politics, archaeology, marine environment, remote sensing, agriculture, censuses, descriptive statistics, statistics, and demographics, among others, in RDR-GEOs. One example is the Database of Places, Language, Culture, and Environment (D-PLACE). These RDR-GEOs offer research data collection on cultural, linguistic, environmental, and geographical information.

The Engineering Sciences area is represented by (10) repositories that cover topics such as Process Engineering, Materials Science, and Computer Science. The main themes available are meteorology, social aspects of transportation, and tectonics. One example is the Monash University Data Repository on Figshare, which has allowed Australian researchers to store, share, and publish Figshare research data since 2014. Every dataset on Figshare is assigned a Digital Object Identifier.

4.7 What software is used to create Geosciences research data repositories?

Finally, the software landscape shows the use of technologies to create repositories. It seeks to answer the following question: what software is used to create Geosciences research

data repositories? The data collected on re3data.org shows that more than half of the software used (179) to create the repositories is unknown, and (129) are classified as other types of software, i.e., (308) repositories are implemented with a variety of specific software technologies for the use of research data, which in general makes it difficult to identify what is used to create these research data repositories. For example, what APIs and protocols are used to provide data interoperability?

A few repositories (17) of the (325) use the Open Archives Initiative Protocol for Metadata Harvesting. Some examples of repositories created with this software are GeoPlatform (Ckan), DataShare (DSpace), the UCLA Social Science Data Archive (Dataverse), OceanRep GEOMAR Repository (e-Prints), and TAMBORA (Fedora).

So, repositories can be implemented with various software technologies, but generally, they are built with free software platforms. The international community and significant university initiatives support using platforms like DSpace, E-prints, Fedora, Dataverse, and CKAN. These free software platforms were not developed to disseminate research data, but they are also being used for this purpose. This software is relevant because it facilitates access, interoperability, and preservation of digital research data. They are used internationally and were developed to collect, preserve, and disseminate publications. However, they can be used to aggregate any content in digital format. It should be noted that although the international community encourages using this software, it is little used in Geosciences and in creating research data repositories.

The following section presents a discussion of the empirical data presented in the research study.

5 Discussion

This section identifies and discusses some of the key points of the study, as well as some of the limitations concerning the scope of the data collected, biases in the selection of RDRs, and technological limitations in web scraping techniques that may affect the generalization of the methodological results.

According to Lin *et al.,* (2024) RDRs are essential for the management, preservation, and sharing of research data, and have become indispensable resources for Geosciences research communities. They provide a central location for researchers to deposit their data and offer a variety of services that facilitate the discovery, access, and utilization of data. RDRs also play a vital role in promoting data sharing and collaboration and help ensure that research data is preserved for future generations.

RDRs are of great importance for the dissemination of research data, enabling open access, sharing, and long-term preservation, which support open science and international collaboration. The global distribution of RDRs shows a concentration in the USA, Germany, and Canada, with most data available in English. This suggests the necessity to increase data availability in other languages to promote global inclusion and accessibility.

The typology of repositories, including disciplinary, institutional, and multidisciplinary, and the variety of deposited data types (images, raw data, office documents, etc.) demonstrate the heterogeneity and richness of available data, indicating the need for standards and best practices for data curation. Most repositories offer restricted access, which raises questions about the necessity of more open and harmonized access policies.

The diversity of software utilized to develop RDRs and the limited adoption of interoperability standards present significant challenges to data integration and exchange. This underscores the vital importance of promoting the use of free software platforms and standardized protocols. The software landscape for the creation of Geosciences research data repositories is diverse. This makes it challenging to conceptualize a computational standard utilized to generate these repositories. This, in turn, introduces a limitation about the interoperability and long-term preservation of Geosciences research data.

As reported by re3data, over half of the software utilized is unidentified. Consequently, it is challenging to identify common software utilized for the creation of these repositories. In general, repositories are constructed with free software platforms that are supported by the international community and university initiatives. The most utilized platforms include DSpace, E-prints, Fedora, Dataverse, and CKAN. Despite their original purpose, these platforms are increasingly being employed for the dissemination of research data due to their capabilities in facilitating access, interoperability, and the preservation of digital research data. Despite encouragement from the international community, the use of this software in Geosciences remains limited.

The study's reliance on data from re3data presents a limitation, as it depends on the continuous validation of this data to ensure relevance and timeliness. So, another limitation is the predominance of repositories in English-speaking countries, which may result in a lack of representativeness of data from other regions and languages. Furthermore, this may lead to an underestimation of the diversity and richness of global Geosciences data. Furthermore, the heterogeneity of data types and software utilized to generate RDRs complicates data interoperability and standardization, impeding integrated analysis and data reuse across diverse research contexts. Additionally, the prevalence of restricted access and specific licenses can

restrict data use and sharing, limiting the potential impact of RDRs in promoting open and collaborative science.

Another limitation of our study is relating a methodological approach. Due to our dependence on technological tools such as web scraping scripts, there may be compatibility issues or failures in the scripts, which could compromise the data collection and analysis. Additionally, the quality of the data collected depends on the accuracy and completeness of the metadata available on re3data. If the metadata is incomplete or incorrect, subsequent analysis may be at risk. Additionally, the methodology may fail to capture all important contexts or details about each repository. These include institutional policies, actual data usage, and the scientific impact of the repositories.

Furthermore, re3data is a dynamic platform, and repository data can be updated frequently. Consequently, there is a risk that the data collected may be outdated by the time of analysis. Although the methodology provides a structured and detailed approach to the analysis of Geosciences data repositories, it is important to be aware of the limitations when interpreting the results and considering possible improvements or additions to the research process.

Finally, the final considerations of this paper are presented.

6 Conclusion

The dimension of Geosciences research data repositories points to a strong trend in studies and practices adopted by information scientists, especially librarians, and data managing and developing library services based on research data.

Research data repositories play a pivotal role in Geosciences, bridging research data users (researchers) and librarians. These repositories, created as platforms to support research and preserve the knowledge derived from scientific investigation, enable data librarians to provide services and products for accessing and safeguarding research data.

In this sense, this study concludes on the international scenario of research data repositories in Geosciences available at re3data. These repositories are heterogeneous data sources that enable access to and preservation of a wide range of types of research data. They store information/knowledge from diverse areas related to the Exact and Earth Sciences, such as Atmospheric, Geodetic, Geophysical, Geological, and Oceanographic Sciences.

The data stored in these repositories encompass many types, including office documents, scientific and statistical data, images, and raw data. This diverse collection of data reflects the multidisciplinary nature of Geosciences, and the variety of research conducted in this field. Most of these repositories are hosted in North American and European countries. In

all, there are (325) cataloged by re3data as of (March 2024), of which the following characteristics can be summarized:

a) the countries with the most repositories are USA, DEU, CAN, GBR, AUS, and FRA, and the predominant language in research data is English;

b) they are in the following areas of knowledge: Atmospheric Sciences and Oceanography, with the main themes being climate, climate change, geology, hydrology, remote sensing and oceanography;

c) most of them are open access and use some Creative Commons licenses;

d) the software used to develop these repositories is primarily unknown.

Looking ahead, we aim to extend the study's methodology to other areas of knowledge. This could involve extracting metadata from Geosciences research data repositories and analyzing the metadata typologies. By doing so, we hope to encourage the replication of our study's approach and findings in other fields, thereby contributing to the broader understanding of research data repositories.

Finally, although international librarians encourage using software such as Ckan, DSpace, DataVerse, e-Prints, and Fedora, they are rarely used in Geosciences. Although they are widely adopted by government institutions and librarian communities that advocate open access to scientific knowledge, developing research data repositories using these software programs is not widespread.

References

BENIN, Keli Rodrigues do Amaral; HAMANAKA, Raíssa Yuri; GONÇALEZ, Paula Regina Ventura Amorim. Digital open repositories: reliability evaluation based on iso 16363 criteria. **Advanced Notes in Information Science**, [S.L.], v. 2, p. 121-130, 2022. DOI: http://dx.doi.org/10.47909/anis.978-9916-9760-3-6.90. Available at: https://anis.pro-metrics.org/index.php/a/article/view/15. Access at: 02 Jul. 2024.

BREMEN Core Repository. **MARUM**: center for marine environmental sciences. Center for Marine Environmental Sciences. 1994. Available at: https://www.marum.de/en/Research/IODP-Bremen-Core-Repository.html. Access at: 02 Jul. 2024.

CANADA. Government. **Canadian Centre for Climate Modelling and Analysis**. 2024. Available at: https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/modeling-projections-analysis/centre-modelling-analysis.html. Access at: 02 Jul. 2024.

CORNELL UNIVERSITY. Mann Library. **Cornell University Geospatial Information Repository**. 1998. Available at: https://cul-it.github.io/cugir-help/about. Access at: 02 Jul. 2024.

EDINBURGH, University Of. **Edinburgh DataShare**. 2024. Available at: https://datashare.ed.ac.uk/. Access at: 02 Jul. 2024.

FIGSHARE. **Figshare**. 2012. Available at: https://knowledge.figshare.com/about. Access at: 02 Jul. 2024.

KHAN, Aasif Mohammad; LOAN, Fayaz Ahmad; PARRAY, Umer Yousuf; RASHID, Sozia. Global overview of research data repositories: an analysis of re3data registry. **Information Discovery and Delivery**, [S.L.], v. 52, n. 1, p. 53-61, 19 abr. 2023. Emerald. http://dx.doi.org/10.1108/idd-07-2022-0069. Available at: https://www.emerald.com/insight/content/doi/10.1108/IDD-07-2022-0069/full/html. Access at: 02 Jul. 2024.

KINDLING, Maxi; PAMPEL, Heinz; SANDT, Stephanie van de; RÜCKNAGEL, Jessika; VIERKANT, Paul; KLOSKA, Gabriele; WITT, Michael; SCHIRMBACHER, Peter; BERTELMANN, Roland; SCHOLZE, Frank. The Landscape of Research Data Repositories in 2015: a re3data analysis. **D-Lib Magazine**, [S.L.], v. 23, n. 3/4, p. 1-10, mar. 2017. CNRI Acct. http://dx.doi.org/10.1045/march2017-kindling. Available at: https://www.dlib.org/dlib/march17/kindling/03kindling.html. Access at: 02 Jul. 2024.

LIN, Dawei; MCAULIFFE, Matthew; PRUITT, Kim D.; GURURAJ, Anupama; MELCHIOR, Christine; SCHMITT, Charles; WRIGHT, Susan N. Biomedical Data Repository Concepts and Management Principles. **Scientific Data**, [S.L.], v. 11, n. 1, p. 1-10, 13 jun. 2024. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41597-024-03449-z. Available at: https://www.nature.com/articles/s41597-024-03449-z. Access at: 02 Jul. 2024.

NASA. **Challenging Mini-satellite Payload (CHAMP)**. 2010. Available at: https://eospso.nasa.gov/missions/challenging-mini-satellite-payload. Access at: 02 Jul. 2024.

PAMPEL, Heinz; VIERKANT, Paul; SCHOLZE, Frank; BERTELMANN, Roland; KINDLING, Maxi; KLUMP, Jens; GOEBELBECKER, Hans-Jürgen; GUNDLACH, Jens; SCHIRMBACHER, Peter; DIEROLF, Uwe. Making Research Data Repositories Visible: the re3data.org registry. **Plos One**, [S.L.], v. 8, n. 11, p. 1-10, 4 Nov. 2013. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0078080. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078080. Access at: 02 Jul. 2024.

PAMPEL, Heinz; WEISWEILER, Nina Leonie; STRECKER, Dorothea; WITT, Michael; VIERKANT, Paul; ELGER, Kirsten; BERTELMANN, Roland; BUYS, Matthew; FERGUSON, Lea Maria; KINDLING, Maxi. Re3data – Indexing the Global Research Data Repository Landscape Since 2012. **Scientific Data**, [S.L.], v. 10, n. 1, p. 1-10, 29 ago. 2023. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41597-023-02462-y. Available at: https://www.nature.com/articles/s41597-023-02462-y. Access at: 02 Jul. 2024.

PANDURO, Anthony Fasanando. Technologies applied to information control in organizations: a review. **Decisiontech Review**, [S.L.], v. 3, p. 1-6, 15 jun. 2023. Pro-Metrics. http://dx.doi.org/10.47909/dtr.02. Available at: https://dtr.pro-metrics.org/index.php/d/article/view/2. Access at: 02 Jul. 2024.

PENG, Ge; GROSS, Wendy S.; EDMUNDS, Rorie. Crosswalks among stewardship maturity assessment approaches promoting trustworthy FAIR data and repositories. **Scientific Data**, [S.L.], v. 9, n. 1, p. 1-10, 21 set. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41597-022-01683-x. Available at: https://www.nature.com/articles/s41597-022-01683-x. Access at: 02 Jul. 2024.

POTSDAM, Geoforschungszentrum. **Scientific Drilling Database**. 2024. Available at: http://www.scientificdrilling.org. Access at: 02 Jul. 2024.

RE3DATA. **Registry of Research Data Repositories**. 2012. Available at: http://www.re3data.org/about. Access at: 02 Jul. 2024.

RICE, R.; SOUTHALL, S. **The data librarian's handbook**. London: Facet Publishing, 2016.
SEMELER, Alexandre Ribas; OLIVEIRA, Arthur Longoni; PEREIRA, Fabiana Andrade; MATIQUITE, Policarpo. Python scripts for web scraping metadata from descriptions of the international scenario of research data repositories. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, [S.L.], v. 28, p. 1-10, 4 ago. 2023. Universidade Federal de Santa Catarina (UFSC). http://dx.doi.org/10.5007/1518-2924.2023.e94877. Available at: https://periodicos.ufsc.br/index.php/eb/article/view/94877. Access at: 02 Jul. 2024.

UNIVERSITY OF BREMEN. **PANGAEA**: data publisher for earth & environmental science. Data Publisher for Earth & Environmental Science. 1987. Available at: http://www.pangaea.de. Access at: 02 Jul. 2024.

UZWYSHYN, Ray. Research Data Repositories: the what, when, why, and how. **Computers In Libraries**, [*s. l*], v. 36, n. 3, p. 1-10, 2016. Available at: https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml. Access at: 02 Jul. 2024.

VALLES-CORAL, Miguel; INJANTE, Richard; HERNÁNDEZ-TORRES, Edwin; PINEDO, Lloy; NAVARRO-CABRERA, Jorge Raul; SALAZAR-RAMÍREZ, Luis; CÁRDENAS-GARCÍA, Ángel; HUANCARUNA, Eddy. Agregación de repositorios institucionales para la generación de información del desempeño científico de universidades peruanas. **Iberoamerican Journal Of Science Measurement And Communication**, [S.L.], v. 3, p. 1-10, 18 out. 2023. Pro-Metrics. http://dx.doi.org/10.47909/ijsmc.63. Available at: https://ijsmc.pro-metrics.org/index.php/i/article/view/63. Access at: 02 Jul. 2024.

WHOI. **WHOI Ship Data-Grabber System**. Available at: http://4dgeo.whoi.edu/shipdata/index.html. Access at: 02 Jul. 2024.

WILKINSON, Mark D.; DUMONTIER, Michel; AALBERSBERG, Ijsbrand Jan; APPLETON, Gabrielle; AXTON, Myles; BAAK, Arie; BLOMBERG, Niklas; BOITEN, Jan-Willem; SANTOS, Luiz Bonino da Silva; BOURNE, Philip E. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific Data**, [S.L.], v. 3, n. 1, p. 1-10, 15 mar. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/sdata.2016.18. Available at: https://www.nature.com/articles/sdata201618. Access at: 02 Jul. 2024.